NOTE ON THE MEASUREMENT OF SKILL OF PROBABILITY FORECASTS

DONALD L. JORGENSEN

Office of Forecast Development, U.S. Weather Bureau, Washington, D.C.

[Manuscript received May 22, 1962; revised September 17, 1962]

ABSTRACT

The issuance of forecasts in probabilistic terms introduces the problem of the measurement of the forecasting skill which goes into the preparation of this type of forecast. The applicability of the conventional skill score to the evaluation of probability forecasts is investigated. It is shown that the skill score can be generalized to give the amount of skill involved in the issuance of forecasts at the various probability levels. The conclusions reached throw some light on the use of the skill score when applied to the conventional two-category type of forecast.

1. INTRODUCTION

In recent years the concept has grown that forecasts could be used more effectively if they were expressed in terms of probability of occurrence of the forecast event. A forecast which verifies correctly in all aspects is the exception, and the nature of the forecast problem requires an explicit statement of the probabilities involved in order to gain the greatest use from the forecasts. Thompson [1] has applied the principle of calculated risk to the forecast problem and has shown how the economy of a weather-dependent operation can be improved by taking into account the probability of occurrence of the unfavorable weather event. A relationship of the form

$$P_d=\frac{C}{L} \tag{1}$$

is given by Thompson where P_d is the decision probability level, i.e., the level above which protective measures will need to be taken in order to minimize loss, C is the cost of taking protective measures, and L is the resulting loss if protective measures are not taken and the weather event occurs. Thompson [2], in a study dealing with the deficiencies of categorical forecasts, and Thompson and Brier [3], developed further the relationship between calculated risk and the usefulness of weather forecasts.

As a result of the growing interest in forecasts expressed in probabilistic terms, the measurement of forecasting skill exhibited by this type of forecast becomes of increasing significance. An extension of conventional methods to the measurement of the skill of probabilistic forecasts is presented here.

2. THE MEASUREMENT OF FORECAST SKILL

For a verification index to measure the skill involved in the preparation of a series of forecasts, the element of chance success (or failure) must be removed from consideration. The conventional skill score first proposed by Heidke [4] and a modification and extension of this type of score by Vernon [5] both take into account the chance element with no additional factors. When only two categories are being considered, the deviation score proposed by Vernon is equivalent to the conventional skill score.

The skill score is not to be confused with other types of indices which measure the operational usefulness of weather forecasts. However, to the extent that the basic value of a forecast is dependent upon the skill involved in its preparation, the skill score is of importance in forecast evaluation.

EVALUATION OF SKILL USING TWO PROBABILITY CATEGORIES

A first approach to the verification of probability forecasts in terms of skill can be made by considering primarily a two-category classification. If the probability of occurrence is above a specified value, the forecast will be counted as an "occurrence" forecast and if the probability is below this value it will be counted as a "non-occurrence" forecast. For better evaluation of the data, they may be distributed in a contingency table in the form given in table 1.

In this table the columns and rows headed W and NW represent the occurrence and non-occurrence events, with F_w and F_{nw} representing the total numbers of forecast occurrences and non-occurrences and with O_w and O_{nw} representing the total observed events. The numbers a

TABLE 1.—*The general form of a contingency table for the evaluation of forecasts involving two categories*

		Forecast		
		W	NW	
Observed	W	a	b	O_w
	NW	c	d	O_{nw}
		F_w	F_{nw}	T

and d represent the correct occurrence and non-occurrence forecasts and the numbers b and c represent the corresponding errors. T represents the total number of forecasts.

A skill score of the form

$$S=\frac{C-E_c}{T-E_c} \tag{2}$$

is evaluated where C is the number of correct forecasts, E_c the number of forecasts expected to be correct on chance. From table 1,

$$C=a+d$$

and

$$E_c=F_w\frac{O_w}{T}+F_{nw}\frac{O_{nw}}{T}$$

which make up the numerator of the skill score. The numerator may be rewritten as follows:

$$C-E_c=\left(a-F_w\frac{O_w}{T}\right)+\left(d-F_{nw}\frac{O_{nw}}{T}\right), \tag{3}$$

where the first term on the right deals only with the occurrence forecasts and the second term with the non-occurrence forecasts. It is of interest that the two terms in parentheses are equivalent, thus

$$\left(a-F_w\frac{O_w}{T}\right)=\left(d-F_{nw}\frac{O_{nw}}{T}\right) \tag{4}$$

It may then be concluded that the skill shown by a series of forecasts involving two categories is evenly divided between occurrence and non-occurrence forecasts. As a result, in evaluating probability forecasts which are classified into two categories, the forecasts which are placed in the high probability category carry the same weight in evaluating the skill as those appearing in the low probability category. No distinction is possible between forecasts issued at varying probability levels within the two categories. In order to evaluate this skill, the forecasts must be broken down into additional categories.

EVALUATION OF SKILL USING TEN PROBABILITY CATEGORIES

In order for probability forecasts to be of greatest value, a group of forecasts issued at a given probability level must verify in nearly the same ratio as the expressed probability. This feature of probability forecasts has been termed the *reliability* of the forecasts, with perfect reliability having been attained when forecasts issued for the individual percentage probability categories are observed to verify with the same percentages. With the development of numerical procedures, it is likely that the goal of perfect reliability can be closely approached. In order to simplify the following discussion, it will be assumed that essentially perfect reliability has been attained. A series of probability forecasts can then be distributed in a contingency table made up of a given number of categories, with the "hits" occurring in the same ratio as the average probability for the individual categories. Tabulating probability forecasts in 10 steps, each with a 10 percent range, gives the generalized contingency table shown in table 2.

TABLE 2.—*The general form of a contingency table for the evaluation of forecasts involving 10 categories*

Observed	Forecast: Ratio of a_i/F_i in percent	$\overline{P}_1$ (95)	$\overline{P}_2$ (85)	$\overline{P}_3$ (75)	$\overline{P}_4$ (65)	$\overline{P}_5$ (55)	$\overline{P}_6$ (45)	$\overline{P}_7$ (35)	$\overline{P}_8$ (25)	$\overline{P}_9$ (15)	$\overline{P}_{10}$ (5 percent)	
	W	a_1	a_2	..	a_m	..	..	..	..	..	a_{10}	O_w
	NW	b_1	b_2	..	..	b_n	..	..	..	..	b_{10}	O_{nw}
	Sums	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	T
						decision probability level						

Table 2 shows that $\overline{P}_i$ is the average probability level of the forecasts within the category, F_i is the number of forecasts issued at this level, a_i and b_i are the numbers of occurrences and non-occurrences in the ith column and a_m and b_n are values of a_i and b_i in the columns adjacent to the decision probability level. O_w, O_{nw}, and T are as previously defined. A more comprehensive evaluation of the skill can be obtained by applying the skill score computation technique to the data in a contingency table of this form. Again the skill of the forecasts given in the above contingency table can be expressed by means of the conventional skill score as follows:

$$S=\frac{C-E_c}{T-E_c}=\frac{\sum_1^m a_i+\sum_n^{10} b_i-\sum_1^m F_i\frac{O_w}{T}-\sum_n^{10} F_i\frac{O_{nw}}{T}}{T-\sum_1^m F_i\frac{O_w}{T}-\sum_n^{10} F_i\frac{O_{nw}}{T}} \tag{5}$$

The numerator may be written in the form

$$\sum_1^m\left(a_i-F_i\frac{O_w}{T}\right)+\sum_n^{10}\left(b_i-F_i\frac{O_{nw}}{T}\right) \tag{6}$$

where the first term deals with the forecasts distributed in the columns above the decision probability level and the second term with the forecasts below this level. Since O_w, O_{nw}, and T are constants for any set of data, the terms are functions of the three variables a_i, b_i, and F_i. The contributions to the overall skill of the individual columns depends upon a_i and F_i in the first term and b_i and F_i in the second term of (6). Here again the two terms in (6) are equal to each other as brought out in equation (4).

Although the evaluation of the skill score as given in (5) depends upon the denominator, the *percentage* contribution of the individual columns to the overall skill can be determined from the two factors making up the numerator as given in (6). If the first factor is represented by M_i and the second by N_i, the percentage contribution of the individual columns can be written:

$$\left.\begin{array}{l}\text{Percentage contribution of}\\ \text{the } i\text{th column for values of}\\ i \text{ ranging from 1 through } m\end{array}\right\} = \frac{M_i}{\sum_1^m M_i + \sum_n^{10} N_i} \times 100 \quad (7)$$

$$\left.\begin{array}{l}\text{Percentage contribution of}\\ \text{the } i\text{th column for values of}\\ i \text{ ranging from } n \text{ through } 10\end{array}\right\} = \frac{N_i}{\sum_1^m M_i + \sum_n^{10} N_i} \times 100 \quad (8)$$

The values of (7) and (8) depend upon the distribution over the 10 columns of F_i, a_i, and b_i, and in addition on the value of the decision probability level P_d which affects the denominator of the two expressions. Given a series of forecasts, the percentage contribution to the overall skill of the forecasts at each probability level can be determined.

As can be seen from (6), those columns for which a_i is less than $F_i O_w/T$ for the high probability forecasts and b_i is less than $F_i O_{nw}/T$ for the low probability forecasts contribute a negative factor to the overall skill. This negative contribution results from the fact that the values of a_i and b_i (the numbers of correct forecasts) are less than would be expected to be correct on a chance basis. If the decision probability level has a value equivalent to the climatological expectancy, then a_i and b_i are always equal to or greater than chance expectancy. As the decision probability level is shifted to higher (or lower) values, the values of a_i (or b_i) in the columns involving probabilities between P_d and the climatological expectancy become less than chance expectancy and a negative contribution is introduced into the skill score. However, because of other factors this does not mean that the skill score has its greatest value when the decision probability level is equal to the climatological expectancy. On the other hand, this does indicate that the value of the skill score obtained for a series of probability forecasts depends on the value chosen for the decision probability level. Since a choice of this probability level is inherent in the preparation of a forecast, whether expressed or implied, it then becomes necessary for forecasts to be based on the same decision probability level in order for the calculated skill scores to be comparable.

TABLE 3.—*Distribution of simulated forecasts in the 10 categories on the assumption of uniform distribution of 250 occurrence cases*

Forecast probability	$\overline{P}_1$	$\overline{P}_2$	$\overline{P}_3$	$\overline{P}_4$	$\overline{P}_5$	$\overline{P}_6$	$\overline{P}_7$	$\overline{P}_8$	$\overline{P}_9$	$\overline{P}_{10}$	Total
a	25	25	25	25	25	25	25	25	25	25	250
b	1	4	8	14	20	30	47	75	142	475	816
F	26	29	33	39	45	55	72	100	167	500	1066

3. ILLUSTRATIVE EXAMPLE

The distribution of an actual series of forecasts in the 10 categories depends upon the climatology of the weather event and upon the skill of the forecaster. For illustrative purposes, the not unrealistic assumption will be made that the occurrences have a uniform distribution in the columns (25 cases per column) with the non-occurrences taking on the required value to give the correct percentage in each column. The assumption gives the distribution shown in table 3.

From relationships (7) and (8), the percentage contribution to the overall skill of the forecasts occurring in each column can be computed for various values of the decision probability level P_d. For values P_d of 20, 50, and 80 percent, the percentage contribution of the skill of each column is indicated in table 4.

The assumed distribution of the forecasts results in a climatological expectancy of 23.4 percent for the occurrence events. As seen in table 4, the percentage contribution of those forecasts in the columns falling between the decision probability level and the climatological expectancy is negative. Since it is impossible to conceive of negative skill (less than complete absence of skill), the negative value represents rather a misuse of the available skill. Although it requires the same skill to issue a forecast at the 75 percent level (column 3) regardless of the value of the decision probability level, it is noted that the contribution of the forecasts falling in this column changes from +8.1 percent to −23.2 percent in going from $P_d = 20$ percent to 80 percent. This change in sign results from the fact that the forecasts issued at this probability level

TABLE 4.—*Percentage contribution to the overall skill of the forecasts occurring in each of 10 categories. The value S of the overall skill is given in last column as obtained from equation (2)*

Forecast probability (percent)	$\overline{P}_1$	$\overline{P}_2$	$\overline{P}_3$	$\overline{P}_4$	$\overline{P}_5$	$\overline{P}_6$	$\overline{P}_7$	$\overline{P}_8$	$\overline{P}_9$	$\overline{P}_{10}$	S
P_d=20	8.9	8.5	8.1	7.5	6.8	5.7	3.9	0.1	6.7	43.3	0.46
P_d=50	11.1	10.7	10.1	9.4	8.4	−7.0	−4.8	−0.1	8.2	54.3	0.49
P_d=80	25.4	24.5	−23.2	−21.5	−19.4	−16.2	−11.1	−1.3	18.8	124.1	0.26

$\downarrow P_d$

TABLE 5.—*Percentage contributions to the overall skill of 1 percent of the forecasts occurring in each of the 10 categories.*

Forecast probability (percent	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
P_d=20	3.6	3.0	2.6	2.1	1.6	1.1	0.6	0	0.4	0.9
P_d=50	8.2	6.9	5.9	4.8	3.7	−2.5	−0.8	−0.1	0.9	2.0
P_d=80	10.3	8.6	↓ P_d −7.4	−6.0	−4.6	−3.1	−1.7	−0.1	1.1	3.2

(75 percent) are occurrence forecasts, based on climatology but by definition are non-occurrence forecasts based on the decision probability level. The occurrence of negative skill in the various columns leads to the paradox shown for $P_d = 80$ percent where the forecasts issued at the 5 percent level (column 10) contribute 124 percent of the overall skill. In order to avoid this type of inconsistency, a skill score used to measure the skill of a series of probability forecasts should be based on a standard value of the probability decision level, preferably equal to the climatological expectancy. This is approximately the case for the data in table 4 for $P_d = 20$ percent.

The number of forecasts in the various columns in table 4 ranges from 26 to 500. In order to evaulate the skill required to issue *one* forecast at the different probability levels (assuming perfect reliability) the values in table 4 need to be divided by the number of forecasts entered in each column. A somewhat more useful value can be obtained by determining the skill contributed by 1 percent of the forecasts. In table 5, percentage contributions of 1 percent of the forecasts are given for the several values of P_d. As shown by the data in table 5, approximately twice as much skill is required to issue a forecast at the 95-percent level as at the 55-percent level, and about four times as much as at the 5 percent level for the assumed forecast distribution. Less skill is required for probabilities near the climatological expectancy and, as would be expected, no skill is required at the 25-percent level which is close to the climatological expectancy.

4. CONCLUSIONS

Although the example given is based on an assumed series of forecasts, the general aspects and conclusions are valid for an actual series. These conclusions may be summarized as follows:

1. Assuming "perfect reliability," the percentage contribution to the overall skill of the forecasts issued at each probability level can be derived. This percentage depends upon the distribution of "hits" and "misses" and the number of forecasts issued at the various probability levels.

2. Negative skill is contributed by forecasts issued at the probability levels between the climatological expectancy and the decision probability level and the value of the computed skill varies with the choice of this level. As a result, it is suggested that a skill score used for the evaluation of probability forecasts be based on the decision probability level equal to the climatological expectancy.

3. The skill contributed by the forecasts issued above the decision probability level is equivalent to that contributed by those issued below this level.

4. Although not stated explicitly, the factors involved in the measurement of skill of probability forecasts also enter into the measurement of skill of the usual type of two-category forecasts. In this latter case, the evaluation of the skill depends upon the assumed decision probability level which in many instances is not definitely stated.

REFERENCES

1. J. C. Thompson, "A Numerical Method for Forecasting Rainfall in the Los Angeles Area," *Monthly Weather Review*, vol. 78, No. 7, July 1950, pp. 113–124.
2. J. C. Thompson, "On the Operational Deficiencies in Categorical Weather Forecasts," *Bulletin of the American Meteorological Society*, vol. 33, No. 6, June 1952, pp. 223–226.
3. J. C. Thompson and G. W. Brier, "The Economic Utility of Weather Forecasts," *Monthly Weather Review*, vol. 38, No. 11, Nov. 1955, pp. 249–254.
4. P. Heidke, "Berechnung des Erfolges und der Gute der Windstarkevorhersagen im Sturmwarningsdienst," *Geografiska Annaler*, vol. 8, No. 4, 1926, pp. 310–349.
5. E. M. Vernon, "A New Concept of Skill Score for Rating Quantitative Forecasts," *Monthly Weather Review*, vol. 81, No. 10, Oct. 1953, pp. 326–329.